UNDERSTANDING SDMX

Introduction to a training course

Table of Contents

Marco Pellegrino (Eurostat)

What is SDMX

This document provides some background information on the SDMX Initiative, the issues which SDMX addresses and the areas in which SDMX is playing a role today. It also provides a description of a training course on the Basics of SDMX, scheduled for GCC-Stat.

1 SDMX in short

SDMX provides support for things that are important to statisticians and are often difficult for them to achieve, and it enables tools to be developed to provide support for these things. It does this by providing standard, well-designed formats for holding all of the elements involved in the statistical process and linking them all together in a clear model. The result is an approach that maximises the amount of metadata and statistical context information that can be passed through to statistical clients, maximises the possibility of relating statistics from similar or different sources, and allows for the automation of processes that are often difficult and costly to manage otherwise.

SDMX, with its standards for metadata and data representation, provides a basis for the development of common tools that can be used by all statistical organisations to improve their dissemination-related activities.

SDMX offers a wide variety of technical tools, and these – like any tools – produce positive results if used well. There is already an impressive array of excellent tools for presenting and working with SDMX data and metadata. But SDMX is not only a standard for data and metadata exchange; it is also a standard for data discovery using web services. Such web-based services enable data and metadata query, visualisation, automated database creation and data loading, metadata query and retrieval from a metadata repository, the linking of data and metadata. Tools exist today that enable the querying of a database, or even file of data as if it is a database, and creation of tables, charts, and graphs from the results of the query. Tools exist to create a database directly from a Data Structure, load the Data Set, and respond to queries for data. Tools also exist to create and maintain SDMX structural metadata and to share this with others.

SDMX also offers a set of guidelines regarding the application of harmonised statistical concepts and codes to data sets, and how these can be represented. Other guidelines address the classification of statistical data and domains, and the harmonization of relevant terminology. Additionally, SDMX represents a framework for the process of harmonization within domains.

In short, SDMX is a model for describing statistical data and metadata, and formats for presenting them with almost unlimited capability for linking the information and adding value in terms of adding quality information, explanatory material, and other forms of aids to interpretation.

SDMX organises and classifies statistical metadata and data to support search and discovery, to link related material, and to link all data to its metadata and to its sources and providers.

As it structures the statistical material in well-defined standard formats, SDMX provides an excellent basis for automated tools to operate on data and metadata – to populate

websites, to provide varied and dynamic data presentations, to resolve queries and find specific details, or to convert material into other formats such as those used by spreadsheets and common statistical tools. There is no better way to support a powerful and flexible dissemination environment for statistical data than to organise that data according to the SDMX Information Model.

2 Why should you be interested in SDMX?

Why should statisticians and National Statistical Offices be interested in SDMX? What does SDMX offer that makes it useful? How will SDMX help statisticians in doing their day-to-day work of producing statistics that are useful and meaningful to clients?

Statisticians need to describe, disseminate, share, relate, and manage their statistical information and processes. They need to publish statistics in ways that convey the full meaning and context to clients, that make quality standards and qualifications clear, and that allow the statistics to be compared and related to other available information. And they would like these products to be produced easily and flexibly with a minimum technical effort. SDMX is focused directly on these requirements. SDMX stands for "Statistical Data and Metadata eXchange" but really SDMX is about much more than pure exchange.

To facilitate data and metadata exchange, one needs:

- a rich and mature model for describing statistical data and metadata;
- comprehensive formats for representing data and metadata, e.g. table structures, related sets of statistical data, and the various quality and other attributes that may be applied to the data;
- well-defined way of making this information available for sharing and access by interested parties;
- tools (operating with the model and the formats/standards) to enable integration, simplify the presentation process, facilitate comparisons of metadata and data, and support collaboration around the data.

This is what SDMX provides and enables. SDMX is a good model for statistical presentation and dissemination activities; the various formats and structures it uses are effective and fit-for-purpose, and an increasing number and tools to work with SDMX-formatted data and metadata are available or becoming available.

Moreover, SDMX has been embraced by the official statistics community – the various national and international statistical organisations that are committed to extracting the maximum value out of statistical information – and its evolution will continue to be driven by these needs. There is a thriving community from these national and international organisations that meets in regular meetings and workshops and in online forums to share ideas, approaches, tools, and requirements.

As an example of a possible use case for SDMX, one could envisage an organisation that:

- produces and publishes statistics accessed by important national clients and by international organisations;
- receives questions both about the origin, production, and quality of the statistics and about comparisons with similar statistics from other countries;
- wishes to make available graphical and flexible presentations of the statistics;

- wishes to automate the process of meeting these requirements to the greatest extent possible to reduce ongoing costs and effort.

SDMX is designed to meet these requirements. It is a standard for machine-processable metadata that carries clear definitions and quality and commentary information, and is an ideal enabler for tools that support self-service and flexibility in presentation.

3 Background: origin and purpose of SDMX

The Statistical Data and Metadata eXchange (SDMX) initiative was launched in 2001 by seven organisations working on statistics at the international level: the Bank for International Settlements (BIS), the European Central Bank (ECB), Eurostat, the International Monetary Fund (IMF), the Organisation for Economic Co-operation and Development (OECD), the United Nations Statistical Division (UNSD) and the World Bank. These seven organisations act as the sponsors of SDMX. They created an initiative with a governing sponsors committee, and a secretariat function to execute the work programme.

The issues can be briefly characterized as follows:

- Statistical collection, processing, and exchange is time-consuming and resource-intensive
- Various international and national organisations have individual approaches for their constituencies
- Uncertainties (in 2001) about how to proceed with new technologies (XML, web services etc.)

In 2001, the SDMX Initiative stated that it would address these issues:

- By focusing on business practices in the field of statistical information
- By identifying more efficient processes for exchange and sharing of data and metadata using modern technology

It was further stated that: “New standards should take advantage of the new web-based technologies and the expertise of those working on the business requirements and IT support for the collection, compilation, and dissemination of statistical information.”

The stated aim of SDMX was to develop and use more efficient processes for exchange and sharing of statistical data and metadata among international organisations and their member countries. To achieve this goal, SDMX provides standard formats for data and metadata, together with content guidelines and an IT architecture for exchange of data and metadata. Organisations are free to make use of whichever elements of SDMX are most appropriate in a given case.

With the Internet and the world-wide web, the electronic exchange and sharing of data has become easier and more common, but the exchange has often taken place in an ad hoc manner using all kinds of formats and non-standard concepts. This creates the need for common standards and guidelines to enable more efficient processes for exchange and sharing of statistical data and metadata. As statistical data exchange takes place continuously, the gains to be realised from adopting common approaches are considerable both for data providers and data users.

SDMX aims to ensure that metadata always come along with the data, making the information immediately understandable and useful. For this reason, the SDMX standards

and guidelines deal with both data and metadata. Common standards and guidelines followed by all players not only help to give easy access to statistical data, wherever these data may be and without demanding prior agreement between two partners, but they also facilitate access to metadata that make the data more comparable, more meaningful and generally more usable.

Thus, the goals of the SDMX initiative were broadly agreed across the sponsoring organisations and within the official statistics community generally. "Official statistics" are the data which is collected and disseminated by a set of governmental and international organizations to provide the factual basis for making policy and supporting research. Some countries have a "national statistical office" (NSO) while others may have several governmental organizations which are charged with collecting statistical data for governmental use. Most countries also have central banks or similar organizations which collect and disseminate financial and economic data. Typically, several national government organizations have a statistical function (ministries of education, justice, labour, etc.)

These national organizations typically report their statistics to a set of supra-national organizations, representing either regions of the globe (examples include Eurostat and the European Central Bank) or domains (examples include the World Health Organization, the Food and Agriculture Organization, UNESCO, and the World Bank). Many of these organizations belong to the UN, or are treaty organizations. All of these organizations exchange, report, and disseminate data in a chain which can be understood as starting at the lowest level within each country, and resulting in high-level data sets which are "aggregated" as they move through various levels to reach the international level.

The system of official statistics is this network of reported data, according to legal requirements or other types of agreements. There are several important meetings, conferences, and initiatives within this system, so that all organizations adopt similar approaches and techniques, and to coordinate reporting: the Conference of European Statisticians is an important meeting, as is the United Nations Statistical Commission meeting. Ultimately the goal is to measure important phenomenon occurring in the world, and to report the data to policy makers, students, journalists, and other users to help inform their activity. The data is "official" because it comes with the reputation of the world's governments and international institutions behind it.

It is important to understand that there were some firm foundations on which SDMX was built:

1. An existing standard for exchanging statistical time-series data, known as GESMES/TS, was already in use among several of the sponsor organizations and their national-level counterparties. This was not based on modern Web technologies such as XML, but used the older UN/EDIFACT syntax.
2. The work on the "metadata common vocabulary" was based on many years of harmonization work within the community, notably Eurostat's *Concept and Definitions Database (CODED)* and the *OECD Glossary of Terms*.

The formation of the SDMX Initiative can be understood as a recognition by the sponsor organizations that working together to address these issues, and that coordinating business approaches using modern, standards-based technology, was the best way forward. In one sense, SDMX evolved from earlier work, but indicated the increased commitment the sponsors had toward reaching its goals. It also represents a coming-together of efforts around harmonizing statistical content and terminology, and for deploying technology to support statistical processes.

Over time, the work of the SDMX Initiative has expanded, both in terms of content-oriented work products and technical ones.

The SDMX Initiative decided early on to position the content-oriented work and the work on technology and standards in a fashion which made these strains of work separate but complementary. The content-oriented work led to the development of the SDMX Content-Oriented Guidelines, while the technical work resulted in the SDMX Technical Specifications. There were several reasons for taking this approach. It reflected the realization that technical specifications must be very precise and detailed in order to allow for automation of statistical exchanges – the programming of computers relies on having very specific rules about how applications communicate, otherwise the communication fails. The SDMX technical standards in one sense function as exchange protocols for machine-to-machine communications (similar to HTTP, for example, but with a focus on specifically statistical exchanges).

Statistical content and terminology issues are very different – they are the subject to interpretation and analysis by trained statisticians. Thus, the technology specifications formed a basis for supporting work on the content side, but in fact are a very different type of work product. It is easiest to see this in the fact that the SDMX Content-Oriented Guidelines are *guidelines,* to help suggest approaches to people in their statistical work, while the SDMX Technical Specifications are *specifications* - rules for developing conforming computer applications.

Another reason for this separation is that the technical specifications and content guidelines were expected to be maintained at different rates – once stable, technical specifications tend to be updated less frequently. Also, the reasons for making updates and changes in each area have no dependency between them, so it made sense to separate them. This is reflected in the fact that the technical specifications are submitted and published through the International Standards Organization (ISO), who publishes many IT-related standards in various domains, while the content-oriented guidelines are not submitted to ISO, but are maintained by the SDMX Initiative itself. This allows for updates of the content-oriented guidelines on an on-going basis.

A third reason for the separation of the SDMX Technical Standards and the SDMX Content-Oriented Guidelines is that – because they are a technological foundation for exchanging *any* statistics – the technical specifications are applicable outside the domain of official statistics, while the content-oriented guidelines are specifically designed to be useful within that context (although they might also be useful outside that community, possibly).

This coordinated-but-separate positioning of the two threads of work has proven to be very useful, too, because often statisticians and economists do not have deep expertise in IT, and technologists do not have deep expertise in statistics. SDMX helps to define the point where the two sets of expertise need to coordinate, to effectively use IT within statistical exchanges and processes.

Within the content-oriented work, there is a set of work products, *The Content Oriented Guidelines,* and 5 annexes:

1. Cross-Domain Concepts
2. Cross-Domain Code lists
3. Statistical Subject-Matter Domains
4. Metadata Common Vocabulary

5. SDMX-ML for the Content-Oriented Guidelines (Concepts, Code Lists, Category Scheme)

The first draft of the Content-Oriented Guidelines was released for public review in March 2006, and a consolidated version was released for public review in February 2008. The full release of the Content-Oriented Guidelines, which have been extensively revised to take account of comments received, took place in January 2009.

4 The existing versions of SDMX

The Version 1.0 SDMX standards were approved by the sponsors in September 2004 and accepted as an ISO technical specification (ISO/TS 17369) in April 2005.

In November 2005, the sponsors approved Version 2.0 of the SDMX standards, which are fully compatible with Version 1.0 but in addition provide for the exchange of reference (explanatory) metadata, and include the registry interface specification.

The SDMX Technical Specifications are now in version 2.1, but both version 1.0 and version 2.0 were implemented.

The 1.0 version of the specifications has a relatively limited coverage – a model for data formats and their structures, along with XML and UN/EDIFACT formats for exchanging these. The UN/EDIFACT format was backward-compatible with GESMES/TS; the XML formats were new. There was also some support provided for SDMX-based Web services: an XML query document, and a set of guidelines about the use of other related Web-services standards (SOAP and WSDL).

The 2.0 version of the technical specifications had a greatly-expanded scope. The model was extended to include “reference metadata” as a way of structuring and formatting metadata related to data quality frameworks, methodological metadata, and other types of “footnote” metadata. Thus, XML formats for reference metadata were added. Further, a set of standard interfaces in XML for interactions with a SDMX Registry were added, for cataloguing the location of data and reference metadata across the Internet or within an organization, and for maintaining and retrieving structural metadata.

In version 2.1, many features of 2.0 have been improved, and the Web-services recommendations have been expanded to include a RESTful interface, standard functions, and error messages. Now, it is possible to develop generically interoperable applications based on the SDMX standards. Further, the various XML data formats have been simplified based on implementation experience with version 2.0.

For all types of work products, there have been internal reviews within the SDMX community, and also public review of the guidelines and standards.

SDMX was approved as an ISO technical specification (ISO/TS 17369) in 2005 and became an ISO international standard (ISO/IS 17369) in 2013.

In March 2007, the sponsoring institutions signed a Memorandum of Understanding (MoU), which is intended to set out the arrangements for a durable collaboration by the sponsors on all aspects of SDMX. The MoU explicitly excludes the formation of any legal entity or common budget for SDMX; each sponsoring institution and its member countries will continue to use its existing procedures to agree on arrangements for transmission and publication of statistics.

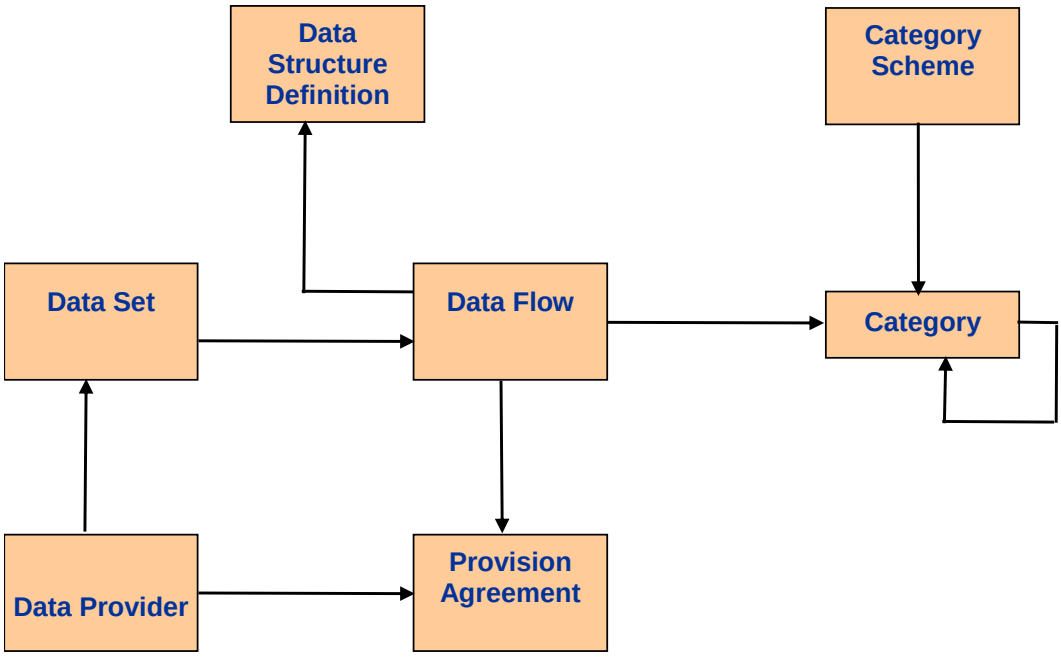
In the conclusions of the 39th Session of the UN Statistical Commission (New York, February 2008), SDMX was recognised and supported as "the preferred standard for exchange and sharing of data and metadata in the global statistical community". This acceptance of SDMX at UN level is a major step forward towards the broader use of SDMX at world-wide level.

5 The SDMX information model

SDMX deals with things statisticians understand well – metadata about concepts, classifications, and table structures, and statistical data related to this metadata. Most of its terminologies are familiar to statisticians although there are a few new terms. Its "artefacts" – the objects it works with and that users must build – are things that statisticians work with on a daily basis: concept definitions, statistical classifications, table structures, data values and data tables, and data attributes like quality information or observation statuses.

The formats SDMX uses to represent artefacts are possibly unfamiliar (XML – ideal for machine automation, JSON, the old GESMES format,…) but statisticians do not need to work at this level: they work with the data and metadata content, while the tools deal with the detail of the format.

The SDMX model for collecting and disseminating data is shown in the figure below.



The key elements of the model are:

Data Flows. Examples of data flows are Quarterly National Accounts statistics and Quarterly (or monthly) Unemployment statistics that many statistical offices produce, along with many other similar regularly-repeated collections. The idea of a data flow is a sequence or series of regularly-repeated data sets coming for different time periods, or from different countries (or both), or being repeated over some other dimension. If we consider National Accounts,

a table typically contains a dimension of measures (Final consumption expenditure – Government and Households, Gross fixed capital formation, etc), a dimension of values (Value, % changed over quarter, % changed over year, etc) and perhaps a Seasonal Adjustment dimension (Original, Seasonally adjusted, Trend). A Data Flow includes dimensions of Time (Q1/2011, Q2/2011, Q3/2011) and perhaps Country (for international organisations such as Eurostat and IMF), for which sub-sets of the data come at different times or from different providers. Thinking of it as a single entity (and holding metadata at the level of this single entity) enables SDMX (and tools working with SDMX) to view all the component data sets as part of this hyper-cube, to understand their commonalities, and to provide very powerful and flexible presentations of the data.

Data Structure Definition. This is the formal definition of the hyper-cube structure for a Data Flow. It defines the hyper-cube in terms of its dimensions, linking them to **Concepts** and **Classifications** (which SDMX calls **Codelists**). It also provides for Attribute information that can provide additional information such as statistical quality, observation status, or other footnote-type information at cell level, at table level, or at intermediate levels.

Data Sets. A Data Set is one of the individual tables that make up a flow. A particular data set might be a National Accounts for France for Q2 2013, for example. A Data Set contains data for some sub-cube of the hyper-cube described by the Data Structure Definition. The Data Flow is made up of its member Data Sets (including those that are still to come in the future).

Category Schemes and **Categories.** These provided indexes for the data flows to allow searching and discovery. Categories can be nested within a Category Scheme and there can be several alternative Category Schemes indexing data flows. A category scheme might be based on a catalogue of publications, or on a dictionary of statistical terms, or on some index of terms in common use amongst a particular set of clients.

Data Providers and **Provision Agreements.** These are the formal mechanisms for capturing agreements to provide data, along with contact details, in a data exchange environment. But they also hold information that is extremely useful in supporting a statistical dissemination website. Importantly, a powerful construct to aid data discovery is the Constraint which holds indexes of the actual content of data sources rather than the theoretical content that is derivable from the Data Structure Definition, thus ensuring that users do not query for data that does not exist. These provision Agreements also provide the basis for automated release calendars and for automated presentation of contact information for queries.

SDMX specifies web services for client programs to query for and retrieve data and metadata directly, it includes mechanisms that support management of complex classifications, classification versions, and related and similar classifications. There are ways to link data flows whose structures contain common dimensions so that tools may harvest this commonality to produce more powerful presentations. There are ways to describe all the elements that go to make up a report or publication to support automatic generation. And there is much more – more that you will discover as you use SDMX.

6 Uptake of SDMX within Domains

SDMX has become very widely used within the world of official statistics, so much so that it is difficult to form a comprehensive list of users. This section attempts to characterize the

current users of SDMX – a group that will likely grow not only in terms of numbers, but also in terms of the breadth of applications. A few possibilities here are suggested at the end of this section.

If we are to look at the most common uses of SDMX, there are two:

1. The use of SDMX as a reporting and collection format, which is prevalent within the central banking community (as a result of the earlier implementation of GESMES/TS, now SDMX-EDI) and among the statistical agencies in Europe (also users of GESMES historically, but implementation is now increasingly driven by such projects as Eurostat's Census Hub);
2. Dissemination of statistical data from websites.

The second application is one which we see in a broad range of institutions, including central banks (ECB and European System of Central Banks, BIS, U.S. Federal Reserve Board and New York Federal Reserve, among others), other sponsoring institutions (IMF, World Bank, OECD, etc.), and national statistical agencies (INEGI in Mexico, Statistics New Zealand, Australian Bureau of Statistics, statistics offices in the European Statistical System etc.)

A less-common but growing use of SDMX is as the basis for data warehouses and other forms of data management. Perhaps the best example of this is the European Central Bank, which has created all of its internal data warehouses around the SDMX Information Model, and has realized many benefits from this. They are by no means the only organization looking at this type of implementation, however – many other organizations are using SDMX to manage not only their statistical data, but also to create metadata repositories, and to integrate their metadata and data.

If we look at which statistical domains have been or are becoming major adopters of SDMX, the list would start with something like this (in no particular order):

- Census and Demography
- National Accounts
- Balance of Payments
- Enterprise Groups
- Labour
- Education
- Food and Agriculture, including fisheries
- Health
- Transport
- Development Indicators
- …

It is easy to see that this is a broad and cross-cutting set of statistical domains – in fact, there are probably very few domains in which SDMX is not being used in some fashion today, and the above list is intended as an indication of the breadth of the uptake..

SDMX was officially endorsed first within the European statistical system, and then by the UN Statistical Commission. These endorsements were powerful incentives for organizations to use SDMX, and the result has been widespread adoption. There are no major competing standards, which has saved the world of statistics from a phenomenon which has slowed the uptake of standards in some other communities.

Additionally, a strong culture of open-source and free tools development has emerged, helping to make the adoption of SDMX easier. This has come both from within the sponsors' community and without, and is supplemented by an increasing number of tools coming from commercial vendors as well.

To learn more about available SDMX tools, the best place is to consult the SDMX website at www.sdmx.org.

The Open Data Foundation hosts the SDMX User Forum in collaboration with the sponsors, providing a place where the community can interact online, and Eurostat's CIRCABC website provides many types of resources, from training videos to student guides. Many organizations offer SDMX in-person training for different levels of users. The best single point of entry is of course the SDMX website itself.

Looking forward, SDMX is increasingly coming into use: Google is using SDMX as a source of data for its Data Explorer; there is now a global registry so that all SDMX data and metadata sources can be easily found; more and more statistical domains are using SDMX for data exchange on a world basis. Furthermore, we see the strong possibility that the world of corporate statistics may realize the utility of having a strong standards basis around the vast amounts of data collected today to support business intelligence applications.

7 How to know more: SDMX Tutorials

SDMX Tutorials have been developed by Eurostat to promote the use and implementation of SDMX standards and guidelines: a set of e-learning videos, each accompanied by a "student book" and a self-test file, is available at http://ec.europa.eu/eurostat/web/sdmx-infospace/trainings-tutorials/tutorials.

This is the list of currently available videos (updates are foreseen):

1. Welcome to SDMX - Why and how SDMX can help you
2. Introduction to SDMX - Outline of SDMX in terms of its history and main components
3. SDMX Information Model - A description of the Information Model and its main objects
4. Data Structure Definition - Explains how data is structured in SDMX
5. Metadata Structure Definition - Explains how SDMX structures reference metadata
6. SDMX-ML Messages - Describes the XML based transmission format supported by SDMX
7. XML based technologies used in SDMX
8. SDMX Architecture for data sharing - How SDMX can facilitate data sharing using different IT architectures

8 How to know more: the SDMX User Guide

Two user guides, one referring to SDMX 2.0 and one referring to SDMX 2.1, are available on the SDMX web site at https://sdmx.org/?page_id=1119.

The new SDMX 2.1 User Guide aims at providing guidance to users of the version 2.1 of the Technical Specification, released in April 2011. As version 2.1 of SDMX contains several

innovative parts (such as web services guidelines, new data messages, new code lists and metadata management) this new release intends to document how the new standard can be used to fulfil the most typical use cases and scenarios for data and metadata exchange.

The principal intention is helping organisations and individuals to determine how best to use SDMX in order to help them to improve the statistical production process. In order to achieve this objective, examples are taken from real implementation scenarios that enable the reader to understand the scope of the SDMX standards and guidelines in terms of the activities required in order to collect, process, and publish statistical data and reference metadata.

Table of contents of the SDMX 2.1 User Guide

Chapter	Content
1. Introduction	Objective, Scope, and Structure of the Guide
2. What is SDMX	Background, sponsors, users, use cases, industry sectors. Brief overview of the technical and content standards, tools, where to find more information and help.
3. Scenario, Use Cases, and Example	This is based on the SDMX Information Model. The chapter relates the Information Model to the real activities of reporting, processing, and dissemination of statistics.
4. Data and Metadata Creation and Reporting	Explanation of the structural components of a Data Structure Definition and a Metadata Structure Definition, and of the Data Set and Metadata Set. How these are used in data and metadata reporting scenarios.
5. Data Bases and SDMX	Explanation of the relationship between the tables in a database and a Data Structure Definition and how the DSD can be used to create these tables. Explanation on how to open a database to SDMX web services.
6. Data and Structure Query	The scope of the data query and the map to the Information Model. Scope of the REST and SOAP queries. Useful tips on what type of queries to support.
7. Metadata Repository and Linking Data to Metadata	Typical requirements for metadata (quality frameworks, linking to disseminated data). Architecture for a metadata repository to enable data and metadata to be combined in a dissemination environment.
8. SDMX Registry	Role of the Registry in statistical data and metadata reporting and dissemination systems. Difference between the content and functions of a Registry and a non-registry based structural metadata repository. Content and role of the Registry in terms of: • Structural metadata maintenance and query • Registration of data and metadata sources • SDMX Registry Services • Web services that can make use of SDMX Registry Services
9. Architecture for an SDMX System	Brings all of the components together in an overall architecture comprising: • Data and metadata persistence and interfaces

Chapter	Content

	• Server side middle tier brokering of requests for data and metadata, data loading, validation, transformation
	• Client side tier of:
	  ○ Structural metadata maintenance
	  ○ Data query and visualization
	  ○ Validation, transformation

10.Community Management	The community may be at the level of an organization or in the context of the wider community of organizations. Topics are:
	• Role of a Global Registry
	  ○ Maintenance agency maintenance
	  ○ Common concepts
	  ○ Community structural metadata
	  ○ Maintenance of Dimension and Attribute roles
	• Hosting of a shared Registry
	• Data provider maintenance